Identity documents classification as an image classification problem

Ronan Sicre, Ahmad Montaser Awal, Teddy Furon

Identity documents classification as an image classification problem

Ronan Sicre, Ahmad Montaser Awal, Teddy Furon

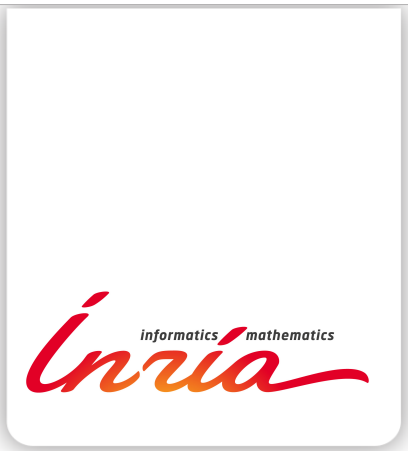
TECHNICAL REPORT

N° 488

2017

Project-Teams Linkmedia

ISSN 0249-0803 ISRN INRIA/RT--488--FR+ENG



Identity documents classification as an image classification problem

Ronan Sicre, Ahmad Montaser Awal, Teddy Furon

Project-Teams Linkmedia

Technical Report n° 488 — 2017 — 11 pages

Abstract: This paper studies the classification of images of identification documents. This problem is critical in various security context where proposed system must offer high performances. We address this challenge as an image classification problem, which has received a large attention from the scientific community. Several image classification methods are evaluated and we report interesting results allowing a better understanding of the specificity of such data. We are especially interested in deep learning approaches, showing good transfer capabilities and high performances.

Key-words: image forensic, image classification, document recognition.

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Classification de documents d'identités vu comme un probleme de classification d'images

Résumé : Ce document porte sur la classification de documents d'identités. Ce probleme est crucial dans de nombreux contextes liés à la sécurité. Nous étudions cette problématique à la manière d'une tâche de classification d'images. Diverses méthodes sont évaluées afin de permettre une meilleure compréhension des spécificités liées à ces données. Nous nous intéressons particulièrement aux approches basées sur l'apprentissage profond.

Mots-clés : sécurité des images, classification d'images, classification de documents

Identity documents classification as an image classification problem

Ronan Sicre, Ahmad Montaser Awal & Teddy Furon

April 27, 2017

This paper studies the classification of images of identification documents. This problem is critical in various security context where proposed system must offer high performances. We address this challenge as an image classification problem, which has received a large attention from the scientific community. Several image classification methods are evaluated and we report interesting results allowing a better understanding of the specificity of such data. We are especially interested in deep learning approaches, showing good transfer capabilities and high performances.

1 Introduction

Identity fraud is a major issue in today's societies with serious consequences. The threats vary from small frauds up to organized crimes and terrorist actions. The work presented in this paper is part of an industrial research project IDFRAud[1] proposing a platform for identity documents verification and analysis. In such a platform, the first step is to collect data, such as document's model from various types of documents as well as emitting countries. In our context, document images can vary from high quality scans to poor quality mobile phone photos and our goal is to identify the type of document and country of origin. This first step is usually followed by document verification, fake detection, documents archiving, etc, which are out of the scope of this paper.

Among image recognition tasks, image classification has been extensively studied over the last decades. These tasks have a large field of applications in search engines, interest object detection, image categorization, etc. The availability of large and/or complex datasets as well as regular international challenges has spurred a large variety of image classification methods. We propose to apply image classification approaches to deal with identity document classification.

Identity documents contain textual and graphical information with a given layout. From such well structured documents, one could expect to extract the document layout as in papers [1, 2] dealing with journal articles, bills, etc. However, the layout is not discriminant when documents from different classes share the same structure or when the documents of the same class do not share the same layout. Other methods are based on text transcription and construct histogram of words appearance combined with standard machine learning classifiers. Unfortunately, such methods are not adapted to our application due to the following difficulties: The document is not localized a priori in the image and background information disturb the classification; text information is difficult to extract before knowing the type of the document. Therefore, we prefer to rely on the graphical content of the identity document and we turn towards image classification techniques to gain robustness and diversity.

[1]This work is part of the IDFRAud project ANR-14-CE28-0012, co-financed by the french DGA: idfraud.fr

Figure 1: Sample images from the databases

2 Previous work

Image classification has received a large attention from the scientific community, *e.g.* see the abundant literature related to the Pascal VOC [3] and ImageNet [4] challenges. A large part of the modern approaches follow the bag-of-word (BOW) approach [5], composed of a 4 step pipeline: 1) extraction of local image features, 2) encoding of local image descriptors, 3) pooling of encoded descriptors into a global image representation, 4) training and classification of global image descriptors for the purpose of object recognition. Local feature points, such as SIFT [6], are widely used due to their description capabilities. Regarding the second step, image encoding, BOW were originally used to encode the feature point's distribution in a global image representation [7, 8]. Fisher vectors and VLAD later showed improvement over the BOW [9, 10]. The third, pooling, step is also shown to provide improvements, and spatial and feature space pooling techniques have been widely investigated [11, 12]. Finally, regarding the last step of the pipeline, discriminative classifiers such as linear Support Vector Machines (SVM) are widely accepted as the reference in terms of classification performance [13].

Recently, the deep CNN approaches have been successfully applied to large-scale image classification datasets, such as ImageNet [4, 14], obtaining state-of-the-art results significantly above Fisher vectors or bag-of-words schemes. These networks have a much deeper structure than standard representations, including several convolutional layers followed by fully connected layers,

resulting in a very large number of parameters that have to be learned from training data. By learning these networks parameters on large image datasets, a structured representation can be extracted at an intermediate to a high-level [15, 16]. Furthermore, Deep CNN representation have been recently combined with VLAD descriptors [17, 18] or Fisher vectors [19, 20].

It is worth mentioning that other approaches have been proposed with the aim of learning a set of discriminative parts to model classes [21, 22]. They are highly effective but costly.

3 Methods

To perform image classification, we first follow the BOW-based pipeline. SIFT keypoints are extracted in either a dense fashion or by interest point detection. Then, these features are encoded with BOW, VLAD or Fisher vectors and are used to classify images with SVM.

Secondly, we study CNN based features, where intermediate transferable representations are computed from pre-trained networks. These image descriptor are similarly given as input to SVM for the classification.

3.1 Bag-of-Words

Assume that the local description output vectors in $\mathbb{R}^d$. The Bag of visual Words aims at encoding local image descriptors based on a partition of the feature space $\mathbb{R}^d$ into regions. This partition is usually achieved by using the k-means algorithm on a training set of feature points. It yields a set $\mathcal{V}$, so called *visual vocabulary*, of k centroids $\{\mathbf{v}_i\}_{i=1}^k$, named *visual words*. The regions are defined as the Voronoi cells of the k centroids. This process is achieved offline and once for all.

The local descriptors of an image $\{\mathbf{x}_t\}_{t=1}^T$ are quantized onto the visual vocabulary $\mathcal{V}$:

$$\mathsf{NN}(\mathbf{x}_t) = \arg\min_{1\leq i\leq k} \|\mathbf{x}_t - \mathbf{v}_i\|. \tag{1}$$

The histogram of frequencies of these mappings becomes the global image description whose size is k.

3.2 Fisher Vectors

Fisher vectors also start from a visual vocabulary $\mathcal{V}$ but used as a Gaussian Mixture Model (GMM). The distribution of the local descriptors is assumed to be a mixture of k Gaussian $\mathcal{N}(\mathbf{v}_i, \mathsf{diag}(\boldsymbol{\sigma}_i^2))$ with weights $\{\omega_i\}$. Covariance matrices are assumed to be diagonal, variances vectors $\{\boldsymbol{\sigma}_i^2\}$ and weights $\{\omega_i\}$ are learned from the training set as well.

Fisher vectors considers the log-likelihood of the local descriptors of the image $\{\mathbf{x}_t\}_{t=1}^T$ w.r.t. to this GMM. They are composed of two gradient calculations of this quantity per Gaussian distribution: The gradient G_μ^X w.r.t. $\mathbf{v}_i$ and the gradient G_σ^X w.r.t. to the variance vector $\boldsymbol{\sigma}_i^2$:

$$G_{\mu,i}^X = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^T \gamma_t(i) \mathsf{diag}(\boldsymbol{\sigma}_i)^{-1}(\mathbf{x}_t - \mathbf{v}_i), \tag{2}$$

$$G_{\sigma,i}^X = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^T \gamma_t(i) [\mathsf{diag}(\boldsymbol{\sigma}_i^2)^{-1}(\mathbf{x}_t - \mathbf{v}_i)^2 - \mathbf{1}_d], \tag{3}$$

where $\gamma_t(i)$ represents the soft assignment term, *i.e.* the probability that descriptor $\mathbf{x}_t$ derives from the i-th Gaussian distribution [9], and where $\mathbf{a}^2$ denotes the vector whose components are

the square of the components of $\mathbf{a}$. The global descriptor is the concatenation of these gradients resulting in a vector of $2kd$ components.

3.3 VLAD

VLAD is a simplified version of Fisher vectors [10] keeping only the difference between the local descriptors and their quantized version onto the visual vocabulary:

$$\mathbf{d}_i = \sum_{\mathbf{x}_t : \mathsf{NN}(\mathbf{x}_t)=i} \mathbf{x}_t - \mathbf{v}_i. \tag{4}$$

The global descriptor $(\mathbf{d}_1^\top, \ldots, \mathbf{d}_k^\top)^\top$ has a size of dk.

A power law, l_2 normalization, and / or PCA reduction are usually performed on Fisher and VLAD [9].

3.4 Convolutional Neural Networks

Deep Convolutional Neural Network [14] are composed of convolutional layers followed by fully connected ones with normalization and/or pooling performed in between layers. There is a large variety of network architectures [23, 24], but a usual choice is 5 convolutional layers followed by 3 fully connected layers. The layers parameters are learned from training data.

The works [15, 25] showed that extracting intermediate layer produces mid-level generic representations, which can be used for various recognition tasks [26] and a wide range of data. In our case, we use a fast network and a very deep network, both trained on ImageNet ILSVRC data. The fast network from [27] is similar to [14], while the deep network stacks more convolutional layers (19 layers in total) with smaller convolutional filters [24].

Following previous works [15, 22, 25], image representations are computed by either taking the output of the fully connected intermediate layers or by performing pooling on the output of the last convolutional layer [25].

4 Experiments

4.1 Datasets

There is no publicly available dataset of identity documents as they hold sensitive and personal information. Three private datasets are provided by our industrial partner. Images are collected using a variety of sources (scan, mobile photos) and no constraint is imposed. Thus, the documents have any dimension, any orientation, and might be surrounded by complex backgrounds. Figure 1 shows examples of such images.

Preliminary experiments is held on a dataset of 9 classes of French documents (FRA), namely *identity card (front), identity card (back), passport (old), passport (new), residence card (old front), residence card (old back), residence card (new front), residence card (new back), driving licence*. A total of 527 samples are divided into train and test, ranging from 26 to 136 images per class. Then, a larger dataset (Extended-FRA or E-FRA) of the same types of documents with a total of 2399 images (86 to 586 per class) is used. The last dataset consists of 446 samples (8 to 110 per class) of 10 Belgian identity documents (BEL), namely *identity card 1 (front), identity card 1 (back), identity card 2 (front), identity card 2 (back), residence card (old front), residence card (old back), residence card (new front), residence card (new back), passport (new), passport (old)*.

Table 1: Evaluation of BOW, VLAD, and Fisher in terms of mAP for detected and dense features, on the FRA dataset.

Encoding	dim.	Detected SIFT	Dense SIFT
BOW 1k	1k	80.7	79.2
BOW 10k	10k	87.0	85.9
VLAD 16	1k	78.5	81.7
VLAD 64	4k	86.6	90.7
VLAD 256	16k	90.1	91.0
Fisher 16	2k	88.9	88.3
Fisher 64	8k	92.8	93.1
Fisher 256	32k	92.7	92.8

Table 2: Performance of several CNN-based features, on the FRA dataset.

Net. layer	dim.	mAP	mean acc.	acc./class
fast fc7	4k	91.1	85.4	85.6
fast fc6	4k	91.7	81.3	81.6
fast c5 - Avg	256	93.2	89.0	90.5
fast c5 - Max	256	92.9	85.7	88.0
vd19 fc7	4k	87.0	81.9	83.2
vd19 fc6	4k	89.4	85.4	86.0
vd19 c5 - Avg	512	89.6	85.4	86.0
vd19 c5 - Max	512	88.3	83.6	82.3

4.2 Results

An extensive evaluation is carried out on the image datasets. Three measures are calculated: mean average precision (mAP), overall accuracy, and averaged accuracy per class.

First, SIFT-based descriptions are evaluated on the FRA dataset, see Table 1. This comprises BOW, VLAD, and Fisher Vector encodings with several visual vocabulary sizes, and from detected or dense SIFT local descriptors. We note that SIFT descriptors are square-rooted and PCA is applied to obtain 64-dimensional vectors. We observe that Fisher Vector performs better than VLAD, which performs better than BOW. This was expected: the more refined is the encoding, the longer is the global descriptor, and the better are performances. Even comparing at similar global descriptor dimension, Fisher Vector is yet the best by a large margin. Note that Fisher Vector does not improve over 64 Gaussians. Secondly, dense local description overall outperforms detected feature except for the case of BOW encoding. This is in agreement with general observations made in computer vision as for classification tasks.

Then, we evaluate CNN-based descriptors on the same FRA dataset, see Table 2. Two architecture are compared: the 'fast' network [27] and the deep 'vd19' network [24]. Descriptors are obtained by extracting the output of the two first fully connected layers (*fc6* and *fc7*), as well as the last convolutional layer (*c5*). Average and max pooling of *c5* are evaluated as well. Surprisingly, the fast network outperforms vd19. Average pooling is also shown to outperform max pooling for convolutional layer and is preferred in the following experiments. Overall *c5* outperforms *fc6*, which outperforms *fc7*. In fact, lower layers (*c5*) encodes lower level and more generic information, which is less sensible to network training data.

Since the CNN feature do not have any rotation invariance mechanism, we propose to enrich the training data collection by adding rotated and flipped images (ending up in 8 distinct descrip-

Table 3: Orientation invariance of CNN features, on the FRA dataset.

Net. layer	mAP	mean acc.	acc./class
fast fc7	92.5	90.5	91.3
fast fc6	92.3	88.1	86.6
fast c5 - Avg	94.1	90.2	90.7
vd19 fc7	89.9	84.8	85.0
vd19 fc6	90.8	87.2	86.5
vd19 c5 - Avg	91.2	88.7	88.8

Table 4: Performance using various combination of the FRA and E-FRA datasets with orientation invariance. Tr Te and E represents Training set of FRA, Testing set of FRA, and E-FRA.

Train/Test Net. layer	mAP	mean acc.	acc./class
Tr/E fast fc7	83.5	81.2	76.7
Tr/E fast fc6	85.6	83.5	78.6
Tr/E fast c5 - Avg	87.8	85.4	83.8
Te/E fast fc7	83.3	83.7	77.8
Te/E fast fc6	84.7	85.8	81.3
Te/E fast c5 - Avg	83.6	86.1	82.1
TrTe/E fast fc7	89.5	86.8	83.9
TrTe/E fast fc6	91.4	89.7	87.6
TrTe/E fast c5 - Avg	90.0	88.5	86.8
TrE/Te fast fc7	97.9	93.6	95.4
TrE/Te fast fc6	99.0	96.3	96.7
TrE/Te fast c5 - Avg	96.6	94.8	95.4
TeE/Tr fast fc7	99.4	96.5	96.4
TeE/Tr fast fc6	99.6	98.0	98.2
TeE/Tr fast c5 - Avg	98.0	94.0	94.4

tors per image), see Table 3. Such process offers a constant improvement for every descriptors.

Further experiments are achieved on the larger E-FRA dataset, see Table 4. Unlike for FRA dataset alone, we observe that *fc6* outperforms *c5*. Unsurprisingly, the more training data the better the performance reaching up to 99% mAP and more than 96% accuracy, when training on E-FRA. More experiment is performed on the BEL dataset, see Table 5. We divide the dataset into three folds, then learn on two third and test on the last one. Scores obtained on all permutations and finally averaged. As for E-FRA, the sixth fully connected layer offers the best performance. Also performances on the BEL dataset are much lower because some classes (residence card (old/front), residence card (old/back), residence card (new/back)) have very few (5 to 12) training samples.

Then, as our application require fast processing of the scanned documents, we evaluate the computation time of SIFT and CNN features extraction, see Table 6. Execution times are computed on a single threaded i7 core 2.6 GHz. Note that image dimensions remained unchanged for SIFT feature, while images are resized to 224×224 for CNN features. CNN features are much faster than SIFT, and keypoints detection is quite slow especially for high-resolution images.

To conclude, CNN generate highly effective compact description, largely outperforming earlier SIFT-based encoding schemes from the classification performance and run-time point of view. Moreover, our evaluation provides insight regarding the amount and balance of data required to reach very high performance.

Table 5: Results obtained on the BEL dataset using 3 folds.

Net. layer	mAP	mean acc.	acc./class
fast fc7	71.7	78.6	64.9
fast fc6	73.8	79.0	66.3
fast c5 - Avg	70.9	77.9	60.0

Table 6: Computation time for detected SIFT, dense SIFT, and CNN features extracted from (224×224) dimensional features on FRA train/test sets.

Features	Det. SIFT	Dense SIFT	CNN
average	54s. / 43s.	5.1s. / 4.9s.	0.2s. / 0.2s.
total	180m. / 240m.	17m. / 27m.	40s. / 53s.

5 Conclusion

This paper addressed the problem of identification documents classification as an image classification task. Several image classification methods are evaluated. We show that CNN features extracted from pre-trained networks can be successfully transferred to produce image descriptors which are fast to compute, compact, and highly performing.

References

[1] V. Eglin and S. Bres, "Document page similarity based on layout visual saliency: application to query by example and document classification," in *ICDAR*, Aug 2003, pp. 1208–1212.

[2] Christian Shin and David Doermann, "Document image retrieval based on layout structural similarity," in *IPCV*, 2006, pp. 606–612.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 Results," .

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Int. Work. on Stat. Learning in Comp. Vision*, 2004.

[6] David Lowe, "Object recognition from local scale-invariant features," *ICCV*, 1999.

[7] L. P. de las Heras, O. R. Terrades, J. Llados, D. Fernandez-Mota, and C. Canero, "Use case visual bag-of-words techniques for camera based identity document classification," in *Int. Conf. on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 721–725.

[8] Jayant Kumar and David Doermann, "Unsupervised classification of structurally similar document images," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1225–1229.

[9] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the European Conference on Computer Vision*. 2010, Springer.

[10] Hervé Jégou, Florent Perronnin, Matthijs Douze, Cordelia Schmid, et al., "Aggregating local image descriptors into compact codes," *Trans. Pattern Analysis and Machine Intelligence*, 2012.

[11] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169–2178.

[12] H Emrah Tasli, Ronan Sicre, Theo Gevers, and A Aydin Alatan, "Geometry-constrained spatial pyramid adaptation for image classification," in *International Conference on Image Processing*, 2014.

[13] Siyuan Chen, Yuan He, Jun Sun, and Satoshi Naoi, "Structured document classification by matching local salient features," in *Pattern Recognition (ICPR), International Conference on*. IEEE, 2012, pp. 653–656.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic, et al., "Learning and transferring mid-level image representations using convolutional neural networks," *Computer Vision and Pattern Recognition*, 2014.

[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database.," in *Advances in Neural Information Processing Systems*, 2014.

[17] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proceedings of the European Conference on Computer Vision*, 2014.

[18] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Lingqiao Liu, Chunhua Shen, Lei Wang, Anton van den Hengel, and Chao Wang, "Encoding high dimensional local features by sparse coding based fisher vectors," in *Advances in Neural Information Processing Systems*, 2014, pp. 1143–1151.

[20] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE CVPR*, 2015, pp. 3828–3836.

[21] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," IEEE *Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[22] Ronan Sicre and Frédéric Jurie, "Discriminative part model for visual recognition," *Computer Vision and Image Understanding*, vol. 141, pp. 28–37, 2015.

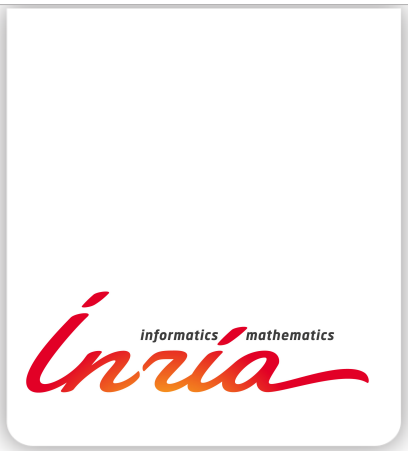
[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[25] Giorgos Tolias, Ronan Sicre, and Hervé Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *ICLR*, 2016.

[26] Ronan Sicre and Hervé Jégou, "Memory vectors for particular object retrieval with multiple queries," in *ICMR*. ACM, 2015, pp. 479–482.

[27] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *BMVC*, 2014.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803